

Einführung der automatischen Indexierung im Österreichischen Verbundkatalog?

Bericht über eine empirische Studie

Otto Oberhauser & Josef Labner

Österr. Bibliothekenverbund & Service GmbH

5. ALEPH – DACH Treffen
Wien, 19. September 2003

ODOK '03
Salzburg, 25. September 2003

Schwerpunkte der Präsentation

- **Allgemeines, Zielsetzung**
- **Methode & Testumgebung**
- **Durchführung der Retrievaltests**
- **Testergebnisse**
- **Resümee & Perspektiven**

Allgemeines, Zielsetzung

Was ist automatische Indexierung ?

Indexieren

Erschließung von Dokumenten durch Hinzufügen von Indextermen (Deskriptoren, Schlagwörter)

Automatisches Indexieren

Alle Verfahren zur automatischen (d.h. nicht-intellektuellen) Dokumentenerschließung
Maschinelle Extraktion (Ermittlung) und/oder Zuordnung von Indextermen zu Dokumenten, mit dem Ziel, eine bessere Basis für das spätere Retrieval herzustellen.

Verfahren für deutschsprachige OPACs

Linguistische, wortbezogene, wörterbuchbasierte Verfahren
Textmaterial: Titel-, Schlagwörter (→ Grundformen, Kompositazerlegung)

Dagegen: statistische Verfahren (Abstract, Volltexte)
regelbasierte (Englisch)
string- u. satzbezogene (Syntakt. Analyse, Kontext)

3

Allgemeines, Zielsetzung

Bisheriger Einsatz in deutschsprachigen Bibliotheken

Projekte / Retrievaltests

- MILOS I (1994/1995): 40.000 Titel / 50 Suchanfragen
 - MILOS II (1995ff): 190.000 Titel / 100 Suchanfragen
 - EKZ-Daten (2000): 47.000 Titel / 30 Suchanfragen
- } *Düsseldorf*
Dipl.Ar. Bonn

- Recall-Steigerung bzw. Anstieg der Anzahl relevanter Treffer
- Precision-Verlust gering
- Reduktion der Null-Treffermengen

- Kascade (1998/1999): gescannte Inhaltsverzeichnisse; Gewichtung

Praktische Anwendung

- ULB Düsseldorf
- Bibliothek, Zentralinstitut für Kunstgeschichte, München
- Friedrich-Ebert-Stiftung, Bonn (Dokumentation)
- Die Deutsche Bibliothek (Retrokatalog 1945ff., in Planung)
- Vorarlberger Landesbibliothek, Bregenz (Inhaltsverzeichnisse)

4

Allgemeines, Zielsetzung

Zielsetzung für den Österr. Bibliothekenverbund

Status Quo (Verbundkatalog / Aleph 500)

- Verbale Sacherschließung / RSWK nur partiell vorhanden (ca. 40%)
- Automatische Indexierung bisher nur durch Anreicherung mit den „Siehe“-Verweisungsformen der SWD

Zielsetzung allgemein

- Prüfung und Bewertung einer zukünftigen Einsatzmöglichkeit eines linguistischen Verfahrens zur weiteren Anreicherung des Standard-Suchvokabulars mit zusätzlichen Suchbegriffen (Indexate)

Zielsetzung konkret

- Realsituation der OPAC-Benutzer: „Alle Felder“-Suche (rund 3/4 aller Suchanfragen erfolgen im Basic Index / UB Innsbruck 2002)
- Vergleich der Retrievalergebnisse VOR und NACH Anreicherung des Basic Index durch Indexate (kein Vergleich einzelner Register wie im Fall der Allegro-basierten MILOS Studien)

5

Methode & Testumgebung

Anwendung von MILOS / IDX für den Test

Möglichkeit zum Testen des Systems MILOS / IDX aufgrund persönlichen Kontakts zu Prof. Lepsky (FH Köln)



Leistungsumfang von MILOS / IDX

- **Grundformenreduktion** (Lemmatisierung)
Häuser → Haus
- **Kompositumzerlegung** (Dekomposition)
Haustür → Haus, Tür
- **Wortableitungen** Adjektiv → Substantiv (Derivationen)
französisch → Frankreich
- **Mehrworterkennung** (bis zu 5 Begriffen)
„Information und Dokumentation“
- **Wortbindestrich-Tilgung**
Geistes- und Kulturgeschichte → Geistesgeschichte
- Einbindung der SWD für **Wortrelationierungen** (auch für Stichwörter!)
Synonyme, Ober- und Unterbegriffe
- *Mehrsprachigkeit (Deutsch, Englisch, Französisch)*
libraries → library (aber nicht Bibliothek)

6

Methode & Testumgebung

Stichprobe für den Retrievaltest

Stichprobenverfahren

Grundgesamtheit (Verbundkatalog):	ca. 3,6 Mio. Datensätze
Angestrebter Stichprobenumfang:	ca. 100.000
Erzeugte Zufallszahlen (Systemnummern):	110.000
davon doppelt u. ausgeschieden:	1.667
Bruttoansatz Stichprobe:	108.333
ohne Löschsätze:	106.904
mit 331, 335, 9xx:	97.460

Laden in einen neuen Aleph-OPAC:	108.333
Exportiert für MILOS / IDX (CD-ROM):	97.460
Automatische Indexierung (Prof. Lepsky):	71.998 (=73,9%)

Die Stichprobe ist hinsichtlich zahlreicher Parameter **ein nahezu exaktes Abbild** des Verbundkataloges

7

Methode & Testumgebung

Testdaten (März 2003)

	Verbundkatalog	Stichprobe
Anzahl absolut	3594154	106904
Hauptsachtitel vorhanden	88,6 %	88,7 %
Zusatz zum HST vorhanden	42,6 %	42,5 %
Beschlagwortet	42,5 %	42,6 %
Sprache Deutsch	60,8 %	61,2 %
Sprache Englisch	21,4 %	21,3 %
Erscheinungsland USA	10,5 %	10,4 %
Erscheinungsjahr 1990	2,9 %	2,9 %
Erscheinungsjahr 2000	3,2 %	3,2 %
Monographien	39,3 %	39,1 %
Zeitschriften & Serien	9,8 %	9,9 %
MAB-Hauptsätze	81,8 %	82,0 %
Kat. Inst. UB Wien	14,2 %	14,0 %

8

Automatische Indexierung der Testdaten

331 Symposium "Baulicher Zivilschutz in Österreich"
335 Rechtsgrundlagen u. Förderungsmöglichkeiten
902 Österreich / Zivilschutz / Kongreß / Wien <1986>

Indexate auf Basis der Titelwörter:

Bau (WA)
Bevölkerungsschutz (SWD-V)
Förderung (ZLK)
Förderungsmöglichkeit (GF)
Grundlage (ZLK)
Möglichkeit (ZLK)
Recht (ZLK)
Rechtsgrundlage (GF)
Schutz (ZLK)
Symposion (SWD)
Symposium (SWD-V)
Ziviler Bevölkerungsschutz (SWD-V)
Zivilschutz (SWD)
baulich (GF)
baulicher Zivilschutz (MWG)
zivil (ZLK)

Indexate auf Basis der Schlagwörter:

Bevölkerungsschutz (SWD-V)
Kongreß (SWD)
Schutz (ZLK)
Ziviler Bevölkerungsschutz (SWD-V)
Zivilschutz (SWD)
zivil (ZLK)
österreichisch (WA)

Positiv fällt z.B. auf, dass die historischen Synonyme zu "Österreich" nicht hinzugefügt wurden (SWD) !



z.B. Ostmark <1938-1945>
Cisleithanien
Habsburgerreich

9

Automatische Indexierung: Fehlermöglichkeiten

331 Reihe österreichischer Autoren

Indexate auf Basis der Titelwörter:

Autor (GF)
Progression <Mathematik> (SWD-V) 
Reihe (SWD)
Unendliche Reihe (SWD-V) 
Verfasser (SWD-V)
österreichischer Autor (MWG)

Jedoch:

Der Anteil solcher und ähnlicher Fehlindexierungen erwies sich in den Testdaten als eher gering !

331 Lise Meitner an Otto Hahn

335 Briefe aus den Jahren 1912 bis 1924 ; Edition und Kommentierung

Indexate auf Basis der Titelwörter:

Brief (GF)
Briefschreiben (SWD-V)
Briefverkehr (SWD-V)
Briefwechsel <Brief> (SWD-V)
Edition (SWD)
Edition und Kommentierung (MWG)

Editionstechnik (SWD-V)
Editionstätigkeit (SWD-V)
Hahn (SWD)
Haushahn (SWD-V) 
Jahr (GF)
...

10

Methode & Testumgebung

Test-OPAC

Nach Durchführung der automatischen Indexierung (Mai 2003):
→ Hinzufügen der neuen Indexate zum Testdatenbestand (Nachladen)

Aufbau von zwei unterschiedlichen Basic Indexes (Stichwörter):

- Alle Felder (so wie bisher)
- Alle Felder mit Indexaten

OPAC für Retrievaltest Automatische Indexierung
[Neustart](#) | [Katalogauswahl](#) | [Optionen](#) |
[Suchen](#) | [Ergebnisliste](#) | [Suchgeschichte](#) | [Bücherkorb](#)

Schnellsuche in allen Feldern

Schnellsuche in allen Feldern mit Indexaten

- Reduzierte OPAC-Funktionalitäten
- Ergebnisliste (Kurztitel) erweitert auf Maximum (99)

11

Durchführung der Retrievaltests

Suchfragen

MILOS II: 100 Suchfragen

z.B.:

Hemmung von Enzymen
Umweltökonomie/ökonomik
Ergonomie
Alternative Energiequellen
Zimmerpflanzen
Strassenkinder
Selbstbewusstsein stärken
Denken und Lernen
Medizin im Dritten Reich
...

Aufbereitung für die „Alle Felder-Suche“ im Aleph-OPAC (Juni 2003)

z.B.:

hemmung AND (enzym OR enzyme)
umweltökonomie OR umweltökonomik
ergonomie
alternative AND (energiequelle OR energiequellen)
zimmerpflanze OR zimmerpflanzen
strassenkind OR strassenkinder
selbstbewusstsein AND (stärken OR stärkung)
denken AND lernen
medizin AND drittes AND reich
...

Anmerkungen

- 99 von 100 Fragen übernommen (Ausnahme: Internet → Intranet)
- Aufbereitung simuliert NICHT ein konkretes Benutzerverhalten, sondern subsumiert mögliche Eingabevarianten, die von „Durchschnittsbenutzern“ erwartet werden können (Einzahl, Mehrzahl)

12

Durchführung der Retrievaltests

Durchführung der Retrievaltests

Retrievaltest 1 (Juni 2003)

- Arbeitsökonomie: Aufteilung auf 12 Mitarbeiter (v. a. aus der Verbundzentrale); Ablauf jedoch rascher als erwartet !
- Pro Person je 8-9 Suchfragen im Basic Index OHNE bzw. MIT Indexaten (nach genauer Vorgabe)
- Protokollierung der Trefferzahl insgesamt sowie der relevanten Treffer
- Keine Recall-Bestimmung (Grösse der Stichprobe !)
- Precision-Bestimmung - "Weit gefasstes" Relevanzkriterium:
Jeder Treffer, bei dem ein Interesse am Originaldokument vermutet werden konnte (MILOS II)
- Redaktionelle Bearbeitung zur Bereinigung einzelner Inkonsistenzen

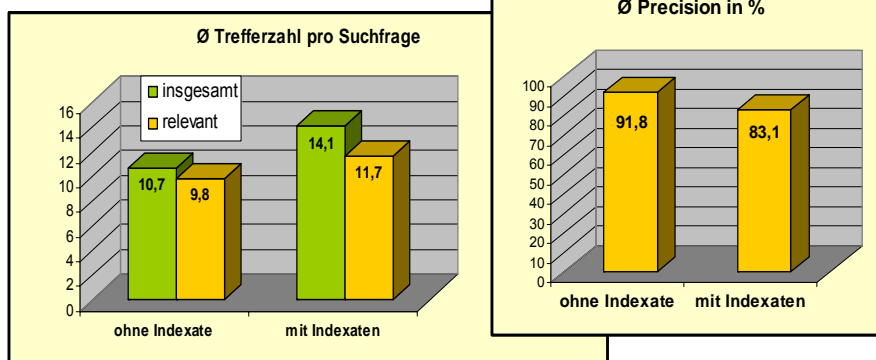
Retrievaltest 2 (Juli 2003)

- Durchgeführt durch die Autoren
- Differenzierung der Resultate hinsichtlich beschlagworteter und nicht-beschlagworteter Datensätze

13

Testergebnisse

Resultate Retrievaltest 1



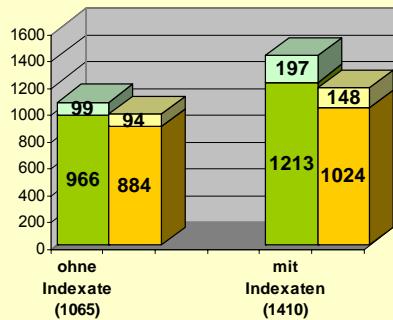
- Der Trefferzuwachs beträgt pro Suchfrage durchschnittlich **3,45** (= 32,4%)
- Der Zuwachs an relevanten Treffern beträgt durchschnittlich **1,94** pro Suchfrage
- Die Relevanz sinkt von ursprünglich **91,8 %** auf **83,1 %**
- Die Zahl der Nulltreffer-Resultate sinkt von **16** auf **11**

14

Testergebnisse

Resultate Retrievaltest 2

**Trefferanzahl pro Suchfrage,
nach beschlagworteten und nicht
beschlagworteten Datensätzen**



Stichprobe: 42,6 % mit SWW
Ohne Indexate: 90,7 % mit SWW
Mit Indexaten: 86,0 % mit SWW

1. Die verwendeten Suchbegriffe führen ursprünglich vor allem zu beschlagworteten Titeln
2. Sofern nicht beschlagwortete Titel gefunden wurden, waren diese kaum weniger relevant
3. Nach der automatischen Indexierung wurden die Suchbegriffe vermehrt auch bei nicht beschlagworteten Titeln gefunden (← Titelwörter)
4. Immerhin sind auch von den 98 hinzugekommenen nicht beschlagworteten Treffern mehr als die Hälfte (54) relevant

15

Testergebnisse

Resultate Retrievaltest 2

	Zuwachs abs.	Prozentwerte
Zuwachs insgesamt	345	32,4 % von 1.065 (Treffer o. Indexate)
davon relevant	194	56,2 % von 345
Beschlagwortet (RSWK)	247	71,6 % von 345
davon relevant	140	56,7 % von 247
Nicht beschlagwortet	98	28,4 % von 345
davon relevant	54	55,1 % von 98

- Der Anteil der beschlagworteten Treffer ist erstaunlich hoch: 71,6 % der 345 hinzugekommenen Titel (Schon vor der automatischen Indexierung waren ja die Synonyme aus der SWD im Basic Index vorhanden waren, sodass zu befürchten war, dass das Verfahren keinen nennenswerten beschlagworteten Titelzuwachs mehr erbringen würde)
- Der Anteil der relevanten Treffer unter den hinzugekommenen nicht beschlagworteten 98 Titeln (28,4 %) ist mit 55,1 % ebenfalls höher als erwartet

16

Testergebnisse

Veranschaulichung der Retrievaltests durch ausgesuchte Beispiele

Retrievaltest 1

Suchanfrage	Eingabe in Alle Felder	ohne Indexate			mit Indexaten		
		Treffer	rel.	% rel.	Treffer	rel.	% rel.
Frau und Beruf	(frau OR frauen) AND beruf	23	18	78,3	48	43	89,6
Geschichte des Mittelmeerraums	geschichte AND mittelmeerraum	30	22	73,3	151	74	49,0

Retrievaltest 2

ohne Indexate						mit Indexaten					
Treffer	rel.	SWW	rel.	oSWW	rel.	Treffer	rel.	SWW	rel.	oSWW	rel.
23	18	21	17	2	1	48	43	44	40	4	3
30	22	30	22	0	0	151	74	143	70	8	4

17

Testergebnisse

Beispieldatensatz: (frau OR frauen) AND beruf

331	Frauen im Erwerbsleben	
335	auf der Suche nach alternativen Formen der Arbeitszeitgestaltung	
902	Frauenarbeit	Frau / Arbeit; Frauenerwerbstätigkeit
902	Flexible Arbeitszeit	Arbeitszeit / Flexibilisierung; Arbeitszeitflexibilisierung
907	Frau	Erwachsene Frau; Weib; Weibliche Erwachsene
907	Berufstätigkeit	Erwerbstätigkeit
907	Arbeitszeitgestaltung	---



IDXt	Arbeitszeit; Arbeitszeitgestaltung; Erwerb; Erwerbsleben; Form; Frau; Gestaltung; Leben; Suche; Weib; alternativ
IDXs	Arbeit; Arbeitszeit; Arbeitszeitgestaltung; Beruf; Berufstätigkeit; Erwerbstätigkeit; Frau; Frau Arbeit; Frauenarbeit; Frauenerwerbstätigkeit; Gestaltung; Tätigkeit; Weib; Zeit; berufstätig; flexibel

18

Resümee & Perspektiven

Bewertung der Testergebnisse

- Der Trefferzuwachs war mit durchschnittlich ca. 1/3 nicht sensationell, aber durchaus respektabel
- Von den Ø 3,45 hinzugekommenen Treffern waren durchschnittlich zwei relevant (= positiv zu bewerten)
- Die Nulltreffer-Ergebnisse sanken immerhin um ein Drittel
- Der Zuwachs an relevanten Treffern bei gleichzeitig nur mässig sinkender Precision, sowie sinkender Nulltreffer-Zahl kann im grossen und ganzen als Bestätigung der MILOS II Resultate gewertet werden
- Nach der automatischen Indexierung wurden (erwartungsgemäss) vermehrt auch Datensätze ohne SWW (bei gar nicht so schlechter Relevanz) gefunden; trotzdem war der Anteil der Sätze mit SWW überraschend hoch
- Der Einsatz der automatischen Indexierung zeigt also gerade beim Vorhandensein der verbalen Sacherschliessung Wirkung und kann nicht als vollständiger Ersatz für diese gelten

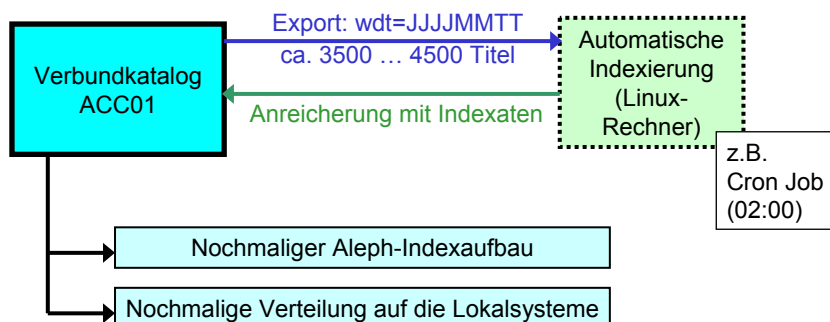
19

Resümee & Perspektiven

Mögliche Implementierung im ÖBV

Erstmalige Anreicherung des Verbundkatalogs und/oder aller lokalen Kataloge
z.B. in sukzessiven Teilschritten, oder
im Ganzen (etwa vor einem Neuaufbau des Verbundkataloges)

Zyklische Anreicherung (z.B. täglich) der neuen bzw. korrigierten Datensätze
inkl. Neuaufbau des Index und Neuverteilung an alle lokalen Kataloge



20

Offene Punkte

- Software (IDX, CAI, ... ?)
- Software-Kosten (IDX, CAI, ... ?)
- Sonstige Kostenfaktoren:
Hardware, laufender Betrieb, ...
- Vollständigkeitsproblem:
Fremdsprachige Datensätze ohne Beschlagwortung (20 – 25%)
- Verteilung im Verbund
Sollen die Lokalsysteme mit Indexaten beliefert werden ?
Zu welchen Bedingungen ?
Wollen die Lokalsysteme überhaupt beliefert werden ?
Falls nicht → unterschiedliche Basic Indices (Multipool-Suche)
- Entscheidung über Einsatz

**Vielen Dank für Ihre
Aufmerksamkeit !**

otto.oberhauser@bibvb.ac.at

josef.labner@bibvb.ac.at